GOOGLE BOOKS, UMI AND OTHER INTRIGUING TRENDS IN DIGITAL PUBLISHING

Joseph G. Wible
Harold A. Miller Library
Hopkins Marine Station
Pacific Grove, CA 93950-3096
wible@stanford.edu

ABSTRACT: For science libraries, journal collections almost always dominate in terms of number of volumes and the percentage taken up by the budget. Therefore, the digitization of journal articles has been a primary focus for many years. Between HighWire Press, commercial publishers, and projects such as JSTOR, this is a rapidly maturing industry. What I want to focus on is the digitization of the book, an area that we have not paid as much attention to in recent years. I will divide my talk into three areas: currently published book, historical book collections, and dissertations.

KEYWORDS: copyright, Google Books, dissertations

Currently Published Books

Administrators often become enamored with the possibility of creating a paperless library, probably because of the false hope that significant money can be saved by taking this approach. For example, when the California State University Monterey Bay was being created in 1994, the original founders had a vision of a university with no "brick and mortar" library. While providing access to over 13,000 journals, they were able to limit the number of journals they subscribe to in paper to 489. But books were another matter. Today they have a 60,000+ volume book collection, and they just broke ground on a new 136,151 square foot library with an initial shelving capacity of 152,000 volumes and a potential shelving capacity of 573,000 volumes.

Today, Stanford is in the early planning stages for building a new engineering library. While in the short term they expect the new library to have a print collection, the hope is that it will be significantly smaller in size than the current library's collection and that eventually the print collection will all but go away.

So the question comes up, how many currently published books are available online today? To determine this, Stanford generated a list of books the library purchased over 18 months between September 2004 and February 2006. The list was limited to books with publication years between 2002 and 2006. We then took a stratified random sample

of 10.2% of the above to create a list of 9271 titles. These titles were then searched in the following sources for full-text books:

Netlibrary
Ebrary
MyiLibrary
Questia
Overdrive
Other (eg. publishers, associations, free-internet)

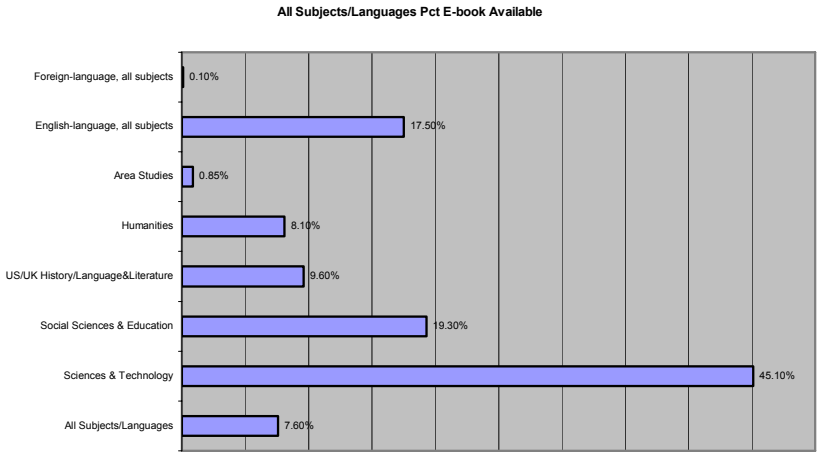
Note that no consideration was given to the quality of the interface, the ability to print, the price for access, etc. The following table shows how the titles fell into broad subject categories. Note that while Stanford libraries as a whole purchased almost 60% non-English books during this time period, for the sciences less than 2% of the 6,720 titles purchased were non-English.

TABLE 1 - Acquisitions for 9 funding clusters 9/1/2004 - 2/28/2006

Fund Cluster	**No. Titles**	**English**	**Non-English**
General Reference	539	510	29
US/UK History/Lang/Lit	7345	7104	241
All other Area & Language	49679	7651	42028
Humanities	18091	9551	8540
Interdisciplinary	928	867	61
Social Sciences & Education	7162	5665	1497
Sciences	**6720**	**6621**	**99**
Media, Reserve, Vickers UG	602	601	1
Totals	91066	38569 (42.4%)	52496 (57.6%)

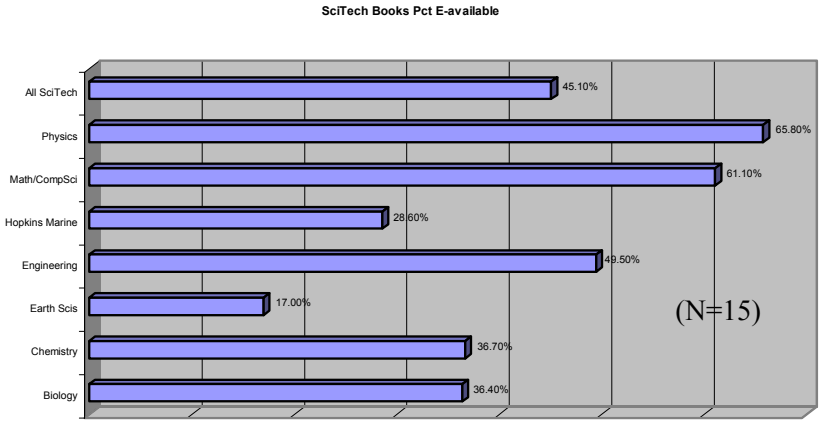
In the chart below you can see the percentage of books available online full-text through any of the sources searched. Note that while the overall percentage for online books was less than 8%, the sciences had the highest percentage with 45%. One contributing factor for the higher percentage is the fact that almost no non-English titles are available online from the sources that were searched. If you eliminate non-English titles, the overall percentage goes up to 18%, but still significantly less than in the sciences.

CHART 1 - E-book availability by broad subject areas



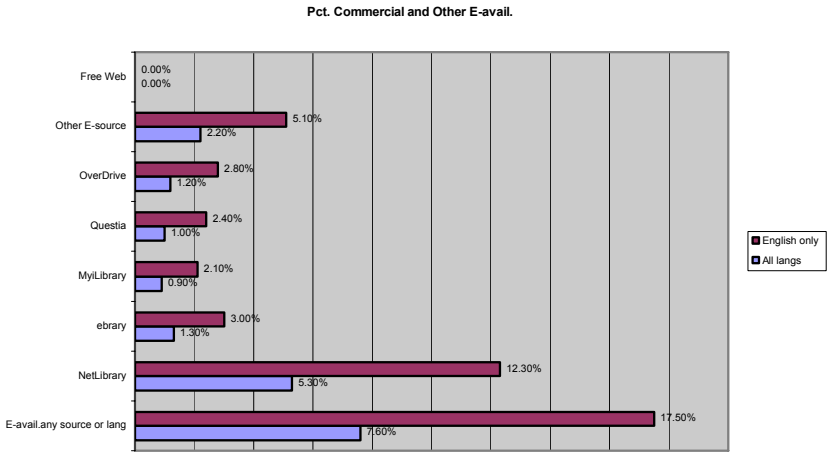
The next chart presents a breakdown among the science disciplines, showing Physics having 66% of its books available online. The marine sciences was less than half of this at 29% available online, but note the very small sample size (N=15).

CHART 2 - E-book availability in 7 broad subject areas of Science and Technology



The next chart shows a breakdown by source of how many books were available online. At 12%, NetLibrary had the highest number of online titles Stanford purchased in print over the 18 month period.

CHART 3 - Overall and relative share of e-book availability among suppliers



*Note: Individual commercial or other e-source percentages are not mutually exclusive, but do reflect, for example, single-source holdings. Hence, the 5.2% difference between

NetLibrary's 12.3% and "E-avail. any source or lang" is comprised of titles singly or multiply held by the other e-sources checked and not held by NetLibrary.

Historical Book Collections

Here my plan was to talk about the Google Books project at Stanford, and particularly at Hopkins Marine Station. When I agreed to give this talk, I had expected my collection to be scanned by Google during the second half of August. Due to a variety of circumstances, this got postponed until September, then October, and now indefinitely. Because of confidentiality agreements Stanford made with Google, I can't say more about this. I also discovered that there were many things I planned to talk about that I couldn't, either because of the confidentiality agreement or because of the lawsuit being filed against Google by publishers.

The first question that usually comes up is why is Stanford participating in the project. What would you do with an offer 1) to digitize every book in your library with no damage to the book 2) to return to you a digital copy for preservation and other purposes, and 3) to present you and the world with a combined word index to millions of books?

From Stanford's perspective this is a great opportunity for digital preservation. After the recent flood that destroyed significant parts of the collection at University of Hawaii, wouldn't it have been nice to have a digital backup copy of all the materials? The other opportunity a comprehensive digital collection presents is the ability to provide enhanced services to the Stanford community. Better navigation tools, citation linking, taxonomic & associative searching and examples of services that could be built on top of the digital archive Google is offering to provide free of charge.

So Stanford made the decision to join Harvard, Oxford, University of Michigan, and the New York Public Library in the Google Books project. Since then University of California and Universidad Complutense de Madrid have joined the project.

As I mentioned earlier, I am prevented from describing some of the process due to the confidentiality agreement Stanford signed. On the other hand, I can present you with information that has been made public. For example, I was told I could not tell you how many books per day are being scanned from the Stanford collection. But because of the Freedom of Information Act and the fact that University of California is a public university, I can tell you that Google is scanning 3,000 books per day from the UC collections. Stanford also learned about how the scanning was being done when it was negotiating with Google, but I am not allowed to tell you how. But if you go to the following URLs you will see the fingers in the scanned pages so you can easy deduce how Google is doing the scanning.

http://books.google.com/books?vid=OCLC03812955&id=1GB1kuY5-pkC&pg=PA3&lpg=PA3

http://books.google.com/books?vid=0sVgqoZH8_0vk2uEA6uPPZ&id=n-28bvRNoroC&pg=RA1-PR1000

http://books.google.com/books?vid=OCLC03812955&id=1GB1kuY5-pkC&pg=PR32

Another thing I can tell you because you can figure it out for yourself is they are scanning EVERYTHING. Try searching Google Books for "36105" and you will get a huge result. This is because the barcodes placed in the back of every Stanford book starts with that number.

I am allowed to tell you that I had no concern about damage that might be done to the collection during the scanning process. I checked with my colleagues on main campus before I agreed to allow Google to scan the Hopkins Marine Station collection. Everyone was satisfied with the care with which the materials were handled. Yes, some materials did get damaged, but these were items that would have been damaged had any library patron picked up the book and tried to read it. If the book spine was too brittle, anyone using it would have to break the spine. Our patrons are often harder on our books that the treatment they received during the scanning process. Stanford views the project as a great way to systematically go through its collection and identify materials that are in need of conservation. As books are pulled for scanning, suspect items are tested to see if the pages are brittle. If they are, they are put aside for the preservation department to treat.

Copyright

I am sure you are all well aware that the publishers are complaining vehemently about the Google Books project. They are also taking Google to court with the claim that it is a violation of copyright law. I, for one, am glad someone with deep pockets like Google, is willing to take on the publishers who continue to push for rights beyond those they are entitled to by law. Libraries often let publishers get away with this because libraries are not willing to fight the battle in court. Even though the law is on their side, defending those rights is still expensive.

It drives me crazy that every time Mickey Mouse is about to go out of copyright, the Disney lobby convinces Congress to change the law to extend copyright coverage additional years. Right now everything published before 1923 is in the public domain. Everything published after 1963 is in copyright and remains in copyright for 70 years after the death of the author. The tricky part is materials published from 1923 through 1963. Materials published during this time period had to have their copyright renewed after 14 years or they became public domain. Only about 15% had their copyright renewed (200,000). The remaining 85% are in public domain. But which are which? How do you figure out whether something is still in copyright when the publisher may have gone out of business? Or was the publisher was absorbed by some other publisher? Even if you contact the publisher, do they have the records to know whether they renewed the copyright, or do they error in their favor and tell you, yes, they still own the copyright?

If you go to this URL: http://www.copyright.gov/orphan/ ,, you can read the report that went to Congress concerning "orphan works". The report recommends that the rights of

the user be protected if the user has practiced due diligence in trying to track down the copyright holder. If they can not locate a legitimate copyright holder and one surfaces later, the report recommends that there be a limit to any remedy that can be sought against the user if they made a reasonable attempt to find the copyright holder and were not successful.

So let's go back to the problem of book published from 1923 through 1963. If any of these books had their copyright renewed, that renewal took place between 1950 and 1992. But there are no electronic records for renewals made from 1950 - 1977. Also, the electronic records for renewal from 1978 - 1992 are very limited in terms of what information they contain. Project Gutenberg scanned and transcribed the printed renewal records which can be found at a series of PDF files arranged by date at URL:

http://onlinebooks.library.upenn.edu/webbin/gutbook/author?name=
United%20States%20Copyright%20Office

Building on this work, Stanford took this data and the electronic records from the copyright office and created a searchable database called "The Determinator" which can be found at URL:

http://collections.stanford.edu/determinator/

Since the copyright records provide very minimal bibliographic information, Stanford is in negotiations with OCLC to see if the records can be matched against its database to provide a richer set of access points to the copyright information. It is also testing the database against manual searches to determine if the use of the database will provide adequately valid results that meet the "due diligence" requirement described in the "orphan works" report. Stanford is vetting this with legal counsel to see if this database will provide a simple and legally safe way of determining whether a book published between 1923 and 1965 is in the public domain.

Putting the legal aspects of Google Books aside, I also find it interesting that publishers are screaming about how this endeavor is taking away their source of income. From my experience it will do just the opposite.

I understand that when National Academy Press started putting the full-text of their new books online for free, it actually increased the sales of their print books. Who wants to read a 400 page book online? Who wants to take the time to print out 400 pages? As long as the book has a reasonable price, most readers would prefer to buy a copy once they have determined that the book is what they want. How do they know they want to buy the book? They know after they have been given an opportunity to read some of it online.

When a faculty or student from Stanford's main campus wants to borrow a book from the Hopkins Marine Station library and there is a full-text version available online, I always

direct them to use the online copy. It cost me money and there is wear and tear on the books when they are shipped back and forth between the two campuses. I can almost guarantee you that the person will respond saying they really want the printed copy anyway. The only time they don't is when they are under a deadline and can't afford to wait for the print copy to be shipped.

I also can attest to the fact that I have purchased books for the Hopkins library as a direct result of the availability of Google Books. Every so often I go into Google Books and search for the phrase "Hopkins Marine Station". Each time I find more books that have information about Hopkins that I was not aware of before because Hopkins wasn't the primary focus of the book. It may have only been a chapter or even just a paragraph mentioning Hopkins. I almost always buy copies of these books to add to my library's collection. These are book sales that would not have taken place without Google Books.

There are also cases where I have been begging publishers to reprint a book that is no longer available. They rarely believe it is fiscally advantageous to do so. Shouldn't they be working with Google to provide a print-on-demand service which would provide the publisher with a new revenue stream?

Publishers are being short sighted and need to start thinking outside the box.

Dissertations
ProQuest (UMI) has been aggressively moving toward digital submission of dissertations. Last year 15% of all dissertations were submitted electronically. This year it has doubled to 30% and an additional 25 schools are in the queue to switch to digital submission. Currently ProQuest has 1.9 million dissertations in microfilm and 800,000 as PDFs. You can check out the online submission process and use the form by going to URL: http://dissertations.umi.com/

Unfortunately Stanford is on the trailing edge in this area. We still submit our dissertations in print. I have been lobbying with the Registrar to change this practice. The reason I feel this is important is because color is now heavily used in many science dissertations. If the dissertation is submitted as a PDF, it will have color. If it is submitted in print, Proquest will scan it to make a PDF, but is only scans in black & white. They have no plans to scan in color because the files created by scanning in color are too large. PDFs created directly from Microsoft Word do not have this size problem. If someone asks to borrow a dissertation from the Hopkins library, I usually would direct them to purchase a copy from ProQuest if the shipping was going to cost more than the purchase price. But many dissertations being produce by today's marine science students are totally useless if color is lost. So I feel obligated to ship copies since there is no alternative. What if my copy gets lost in shipping? The "backup" copy at ProQuest is not an acceptable backup since it does not have color.

Conclusion

We are still in the early to middle stages of migrating from a print environment to a digital environment when it comes to books. The implications of the switch in terms of the traditional economic model and the existing copyright law are major, which makes life interesting for the practicing librarian.